”LOW SUPERVISION” DEEP CLUSTER CHANGE DETECTION (CDCLUSTER) ON REMOTE SENSING RGB DATA: TOWARDS THE UNSUPERVISING CLUSTERING FRAMEWORK

Guglielmo Fernandez Garcia, Iris de Gelis, Thomas Corpetti, Sébastien Lefèvre, Arnaud Le Bris

▶ **To cite this version:**

Guglielmo Fernandez Garcia, Iris de Gelis, Thomas Corpetti, Sébastien Lefèvre, Arnaud Le Bris. ”LOW SUPERVISION” DEEP CLUSTER CHANGE DETECTION (CDCLUSTER) ON REMOTE SENSING RGB DATA: TOWARDS THE UNSUPERVISING CLUSTERING FRAMEWORK. IGARSS 2023 - 2023 IEEE International Geoscience and Remote Sensing Symposium, Jul 2023, Pasadena, United States. pp.6656-6659, 10.1109/IGARSS52108.2023.10282898 . hal-04309802

HAL Id: hal-04309802

https://hal.science/hal-04309802v1

Submitted on 27 Nov 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

"LOW SUPERVISION" DEEP CLUSTER CHANGE DETECTION (CDCLUSTER) ON REMOTE SENSING RGB DATA: TOWARDS THE UNSUPERVISING CLUSTERING FRAMEWORK

Guglielmo Fernandez Garcia [1], Iris de Gelis[2], Thomas Corpetti[1], Sébastien Lefèvre [3] and Arnaud Le Bris[4]

[1] CNRS, LETG UMR 6554 Rennes, F-35000, France
[2] Magellium, F-31000 Toulouse, France
[3] Université Bretagne Sud, IRISA UMR 6074, F-56000 Vannes, France
[4] Univ. Paris-Est, LASTIG STRUDEL, IGN, ENSG, F-94160 Saint-Mandé, France

ABSTRACT

This paper is concerned with the change detection issue in remote sensing images. This problem is not trivial since the notion of change depends on the application. Moreover, classical supervised deep learning methods have to deal with the limited amount of labelled data available. Based on existing deep learning techniques that exploit unsupervised clustering to assign labels to entire images, we adapt them to the change detection problem by using siamese backbones and extracting pixel-wise results. As fully unsupervised experiments lead to unstable results, we suggest "low supervision" strategy composed of a *warm-up* stage with few labeled data able to drive the following unsupervised learning through reliable solutions. Preliminary experiments show reliable change maps.

Index Terms— Change Detection, Semi-supervised Learning, Unsupervised Learning, Clustering, Multitemporal images, Deep learning.

1. INTRODUCTION

The availability of advanced remote sensing (RS) imagery has paved the way for a new era of environmental studies [1] and Earth Observation (EO) studies. In addition to satellite images [2], researchers now have access to unmanned aerial vehicle (UAV) or aerial surveys [3], as well as historical archives [4]. These resources offer valuable opportunities for investigating the dynamics of various areas, such as land resource planning, disaster monitoring, and urban expansion, by analyzing dense time series data. A critical aspect in these studies is Change Detection (CD) [5], which involves understanding the progression of events over time, specifically the analysis of bi-temporal images, to output change maps that identify, and eventually classify, change areas.

Historically, CD has been tackled as an unsupervised problem based on the analysis of differences between the two input images, as in the well-known CVA method [6, 7]. These methods, however, are often based on a pixel-wise comparison of images. As RS imagery embed high spatial correlation, such approaches may limit their capability to obtain further information on the nature of the change. More recently, deep learning, specifically Convolutional Neural Networks (CNNs), have captured researchers' interest in developing novel CD methods thanks to their capacity to capture spatial context. However, CD *via* CNNs remains challenging when dealing with high-resolution remote sensing data. One of the primary difficulties is the limited availability of labelled data necessary for training. Historical aerial image archives exemplify this issue: although they provide rich time series data with high spatial resolutions (sub-meter) and diverse spectral channels (RGB, infrared, panchromatic, etc.), these parameters vary throughout the time series. Consequently, this heterogeneity results in a notable scarcity of labelled data, necessitating larger-than-usual resources for the labelling process.

To cope this difficulty, unsupervised deep learning approaches have been recently developed with remarkable results (for examples see [8, 9, 10]). Nevertheless, depending on applications, these methods are often limited with respect to the diversity and complexity of data. Examples of open problems [11] can be related to multisensor data, multiscale changes, classification or finally the inherent problem of defining what the change is (e.g. a change in the canopy of vegetation or the construction/destruction of a building).

Recent years have seen the development of datasets oriented towards CD, like LEVIR-CD [12], HRSCD [13] or SECOND [14], creating a common basis for benchmarking methods but also providing data availability that can be exploited for pre-training models. This also enabled the development of novel insights from semi-supervised or supervised learning on the issue [15, 13, 16].

Given all these elements, we propose in this paper the first steps for a novel approach for pixel-wise CD, called "low supervision" strategy, that can be a middle-ground between supervised and unsupervised learning. The approach exploits a few labelled training data to bootstrap unsupervised training with a more significant number of unlabelled training samples. To do so, we took inspiration from similar works in the field of visual feature clustering, primarily DeepCluster [17, 18], in which an unsupervised clustering method is used to generate pseudo-labels from features extracted via a CNN; these pseudo-labels are then used to optimize the whole network. The method presented, entitled CDCluster (see Figure 1), can be seen as an extension of the latter idea to CD: as in classic siamese networks [19, 20], two encoders generate features per date that are subtracted, followed by a decoder and convolutional layers; such features are then clustered and used as pseudo-labels.

2. MATERIALS AND METHODS

In this section, the overall pipeline of our CDCluster is given (Figure 1. Then, a detailed description of the training and validation strategy is provided.

This work has been supported by the ANR HIATUS (ANR-18-CE23-0025), financed by the French National Research Agency.

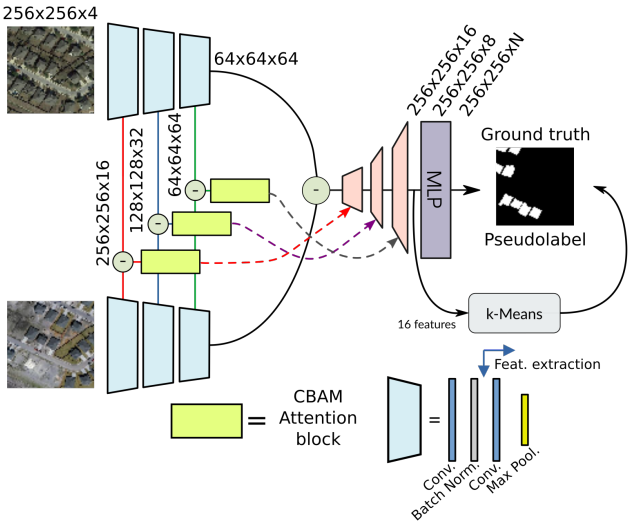

Fig. 1. Schematic representation of CDCluster model.

2.1. Feature extractor

As mentioned, one of the main inspirations behind CDCluster is the capacity of siamese networks[19, 20] to extract a set of features adapted to CD. First, the RGB images are treated with a Sobel filter, and the input vector is formed among these four channels. The idea is to guide the network with realistic solutions. Secondly, the dimensionality of the input vectors (256x256x4) is reduced through two encoders, composed of three blocks, to arrive at two feature spaces of 64x64x64 that are subtracted. Each block consists of a convolution layer, followed by batch normalization, a second convolution layer and a max pooling. A decoder is used to upscale these features up to 256x256x16, the size at which the features are clustered. However, in the unsupervised/"low supervised" settings, there is no guarantee that the extracted features actually contain the right amount of information about the differences between the two inputs to be able to generate pseudo-labels. For this reason at each encoding block, after normalization, features are extracted from each channel and subtracted from each other. Three skip connections, with CBAM attention [21], are used between the encoder's features differences and the decoder. On top of this architecture, two dense layers forming an MLP (Multi-Layer Perceptron), extract the change detection maps. The number of outputs, corresponding to the number of clusters (k), is treated as a hyperparameter (see the following section). During all training, categorical cross-entropy was used as loss and learning rate of 0.001.

2.2. Feature clustering and pseudo-label generator

The second core idea of CDCluster is to assign pseudo-labels via clustering the 256x256x16 extracted features. In this work, we focused on the k-Means implementation present in the FAISS package (that allows fast search on GPU devices). As suggested by Caron et al. [17] and Mustapha et al. [22], it is not the specific choice of clustering algorithm that is crucial, but rather the number of clusters, k, and the "clustering halt", that is the frequency of epochs at which the centroids are calculated. Concerning the number of clusters, it has been shown on similar tasks that k needs to be larger than the number of desired classes to ensure the stability of the model. For this reason, we explored a set of k ranging from 2 up to 500 in preliminary studies. A final value of $k = 20$ has been set. Concerning the clustering halt, no influence has been found after a value of 5 epochs. In order to extract binary change maps from the 20 clusters, the per-pixel maximum value has been taken and binarized. Similarly than DeepCluster, CDCluster is also prone to trivial solutions. Therefore, while training, whenever the population of a cluster drops below a given threshold, the data points assigned to a random above-threshold cluster are split and reassigned to the below-threshold one. At each epoch of clustering, only 75% of the dataset was used, chosen randomly, in order to limit GPU memory.

2.3. Dataset and training strategy

For this work, we decided to use a well-established dataset for CD, namely the LEVIR-CD dataset [12], in order to be able to validate and test the results with standard evaluation metrics.

In an unsupervised setting, in order for CDCluster to be able to discriminate the change zones between the two images, it is necessary that the features generated by the convolutional filters in the first epochs of the training are consistent with regard to the CD issue. Preliminary tests in this setting did not lead to satisfactory results. A first possible solution, not explored here, might be to encode more efficiently the differences between the images, as suggested in [23]. Alternatively, it can be assumed that a small amount of ground truths exists and they can be used to bootstrap the initial signal. For this reason, we divided the training into two phases: a "warm-up phase" and an "unsupervised phase". The former consists in a supervised phase in which only 30% of the dataset is used for training. Since the quantity of True Negative data is not negligible in the dataset, a data sampling based on a minimum quantity (3%) of True Positive pixels has been used, except where noted. Subsequently, in the "unsupervised phase", the model is free to explore the whole LEVIR-CD dataset and the pseudo-labels generated from clustering are used. The term "low supervision" used in this paper, therefore, indicates the sum of the two phases, in which the initial signal given by the random values of the convolutional filters is augmented with the help of a small part of labelled data.

2.4. Metrics and validation strategy

Standard metrics such as Accuracy, F1-score, Mean Squared Error (MSE), Cohen's kappa (κ), Specificity and Confusion Matrix have been used to evaluate the results on the test set. Apart from this, we aimed also to explore and understand under which conditions CDCluster works. It has been pointed out that the quality of the random filters at the beginning of the training must play a fundamental role, and they can be used as a proxy for the final performances of the training [22]. In order to assess this factor for CD, we also used the Initial Alignment (IA) metric, introduced by Mustapha et al.[22], defined as:

$$IA(k, \theta) = NMI_{adj.}(L, P) \quad (1)$$

where k is the number of clusters, θ is the random filters generated for a given seed at the beginning of the training, L is the set of ground truth labels, P is the set of pseudo-labels generated from the starting θ random filters and $NMI_{adj.}$ is the Adjusted Normalized Mutual Information.

3. RESULTS AND DISCUSSION

As a first evaluation, we explored the performances of CDCluster on the test set (Table 1). Remarkably, a high level of Accuracy and

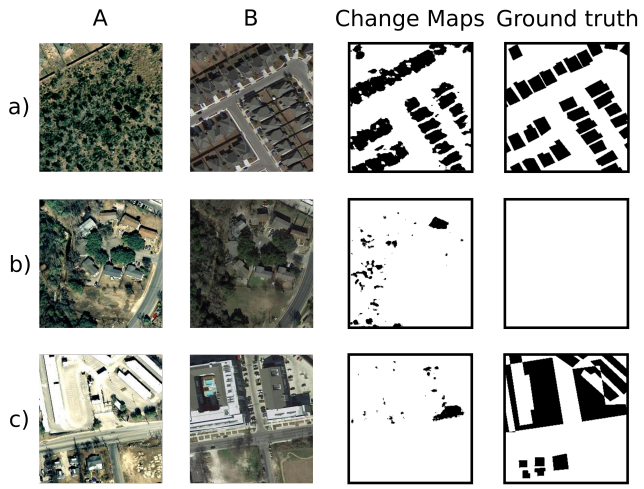

Fig. 2. Examples of predictions: couple of images A and B are shown, along with the predicted Change Map and the ground truth (used only for the calculation of metrics during testing). Several cases are reported as example: a True Positive change maps (a), False Positive (b) and False Negative (c).

Specificity was found. As suggested by the F1-score and the κ, it is expected, however the presence of False Negative detections. The Confusion Matrix further confirms this. These results are in line with the state of the art of the literature [12]. When looking at the graphical results of the detection (Figure 2), it is evident that an absence of constraints in the definition of the contours of the change maps undermines the overall quality of the detection. This is rational, considering that no element is currently provided in the backbone to suggest the definition of more accurate contours. An initial analysis found no correlation between building size and False Negatives. As far as False Positive pixels are concerned, it should be noted that their evaluation is not trivial: the LEVIR-CD dataset is focused on the detection of buildings, and it cannot be excluded that these False Positives are other types of change (e.g. vegetation) not included in the ground truth.

Accuracy	F1	MSE	κ	Specificity	Confusion Matrix
0.98	0.75	0.02	0.74	0.99	$\begin{pmatrix} 0.99 & 0.30 \\ 0.01 & 0.70 \end{pmatrix}$

Table 1. Metrics on the test set for CDCluster with $k = 20$ and 30% of data in the "warm-up" phase.

In order to consolidate this approach, however, it is also necessary to explore under which conditions these results are achievable. As mentioned, k and IA can play a pivotal role in this kind of approach and these hyperparameters depend on the specific model and the dataset. For assessing these factors, we calculated the IA for eight seeds for a range of k from 2 to 500 (Figure 3). First of all, it has to be noted that all values of IA are at least one order of magnitude lower than similar tasks [22], suggesting an explanation on the reason why the "low supervision" approach is necessary. Secondly, after approximately $k = 10$, the average IA is independent of k, supporting the choice of $k = 20$ to balance accuracy and computational costs. Finally, it should be noted that IA with different clusters, but

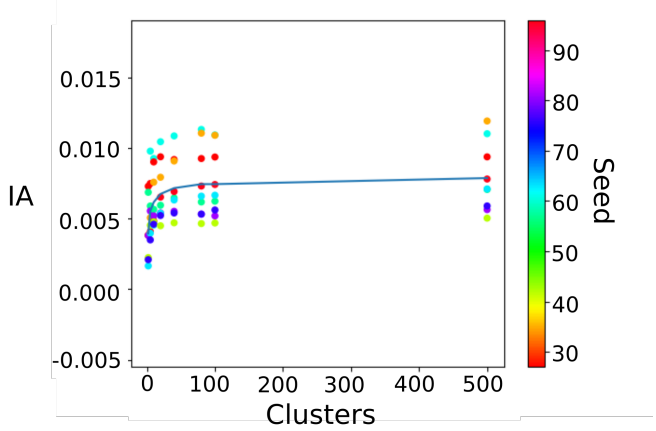

Fig. 3. Initial Alignment (IA) dependency on the number of clusters (k) for different seeds. The blue line is the average of all seeds.

the same seed, trace a consistent trend with only minor differences. While this suggests that, in our particular case, the choice of seed is not crucial, it also support the idea that, given a seed and a k, the value of IA with another k can be used as proxy for preliminary explorations.

These results raise the question of the amount of data needed in the "warm-up phase", as it is desirable for this phase to be as short as possible during the training but, above all, with as little data as possible. For this reason, we tested different amounts of data in the "warm-up phase", randomly sampled from the full LEVIR-CD dataset. As shown in Table 2, no relevant results can be found up to 20% of the train set, corresponding to 1424 couples of images. On the other hand, when applying a sampling only on data with at least 3% of True Positive pixels (as mentioned in Section 2.3), we were able to recover most of the accuracy with only 6% of the dataset, corresponding to 427 couple of images. Even if it requires further experiments, this result consolidates the idea that a "low supervision" approach, i.e., using a minimal amount of labelled data at the beginning of the training, can be explored as a viable idea to initialise unsupervised learning.

Metric	1%	2%	5%	10%	20%	6% (sampled)
Accuracy	0.05	0.05	0.05	0.05	0.84	0.72
MAE	0.94	0.95	0.95	0.95	0.10	0.24

Table 2. Accuracy and MAE with the reduction of data in the "warm-up" phase.

4. CONCLUSIONS

In this work, we explored a novel pathway to perform CD in a setting that mixes supervised and unsupervised learning, which we called "low supervision": based on the presence of a modest portion of ground truth, an initial supervised ("warm-up") phase gives clues to the model as to what kind of changes are being sought, followed by an unsupervised phase. In doing so, we also presented an extension of DeepCluster to the CD, based on the idea of Siamese networks. From this point of view, one of the fundamental questions of this work can be reformulated as follows: under what conditions is it possible to exploit visual features from a neural network to create

pseudo-labels by clustering? Of course, giving an unambiguous and definitive answer is beyond the scope of this preliminary work, but we can already state some elements. For example, it should be noted that CDCluster shows encouraging results for $k = 20$ and 30% of the dataset (with 3% white pixels), indicating that it is indeed possible to obtain valuable information from change features. The low values of IA corroborates the need of an initial bootstrapping. On the other hand, these results lead to the question of which are the ideal conditions for a backbone to effectively encoding change information, question that is shared also in other fields, as 3D Point Clouds CD [24]. In this sense, this paper aims to take a first step towards a more systematic study of how to create architectures with higher IA.

Beyond these considerations, this study shows another interesting possibility. The increase of public datasets for CD can become a good starting point for developing strategy for transfer learning to a specific dataset. Although in this work both training phases involved the same dataset, the prospect to be explored in the future is to perform the unsupervised phase on an unknown dataset (such as the HRSCD dataset). In this regard, it remains to be determined whether the current minimum amount of data to obtain results is sufficient and evaluate the generalization possibilities for features calculated on one dataset to be exported to another.

5. REFERENCES

[1] Roberta Kwok, "Ecology's remote-sensing revolution," *Nature*, vol. 556, pp. 137–138, 04 2018.

[2] J. Yang et al., "The role of satellite remote sensing in climate change studies," *Nature Climate Change*, vol. 3, pp. 875–883, 09 2013.

[3] Alvarez-Vanhard E. et al., "Can uavs fill the gap between in situ surveys and satellites for habitat mapping?," *Remote Sensing of Environment*, vol. 243, pp. 111780, 06 2020.

[4] A. Le Bris et al., "Cnn semantic segmentation to retrieve past land cover out of historical orthoimages and dsm: first experiments," *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. V-2-2020, pp. 1013–1019, 08 2020.

[5] Lazhar Khelifi and Max Mignotte, "Deep learning for change detection in remote sensing images: Comprehensive review and meta-analysis," *IEEE Access*, vol. PP, pp. 1–1, 07 2020.

[6] R.D. Johnson and E.S. Kasischke, "Change vector analysis: A technique for the multispectral monitoring of land cover and condition," *International Journal of Remote Sensing*, vol. 19, no. 3, pp. 411–426, 1998.

[7] F. Bovolo and L. Bruzzone, "A theoretical framework for unsupervised change detection based on change vector analysis in the polar domain," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 45, no. 1, pp. 218–236, 2007.

[8] L. Bergamasco et al., "Unsupervised change detection using convolutional-autoencoder multi-resolution features," *IEEE Transactions on Geoscience and Remote Sensing*, vol. PP, pp. 1–1, 01 2022.

[9] Y. Sun et al., "Structure consistency-based graph for unsupervised change detection with homogeneous and heterogeneous remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. PP, 01 2021.

[10] Luppino L. T. et al., "Code-aligned autoencoders for unsupervised change detection in multimodal remote sensing images," *CoRR*, vol. abs/2004.07011, 2020.

[11] H. Jiang et al., "A survey on deep learning-based change detection from high-resolution remote sensing images," *Remote Sensing*, vol. 14, no. 7, 2022.

[12] Hao Chen and Zhenwei Shi, "A spatial-temporal attention-based method and a new dataset for remote sensing image change detection," *Remote Sensing*, vol. 12, no. 10, 2020.

[13] Caye Daudt R. et al., "Multitask learning for large-scale semantic change detection," *Computer Vision and Image Understanding*, vol. 187, pp. 102783, 2019.

[14] Kunping Y. et al., "Asymmetric siamese networks for semantic change detection," *CoRR*, vol. abs/2010.05687, 2020.

[15] R. Caye Daudt et al., "Fully convolutional siamese networks for change detection," in *2018 25th IEEE International Conference on Image Processing (ICIP)*, 2018, pp. 4063–4067.

[16] C. Sun et al., "Semisanet: A semi-supervised high-resolution remote sensing image change detection model using siamese networks with graph attention," *Remote Sensing*, vol. 14, no. 12, 2022.

[17] M. Caron et al., "Deep clustering for unsupervised learning of visual features," in *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[18] M. Caron et al., "Unsupervised learning of visual features by contrasting cluster assignments," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, Eds. 2020, vol. 33, pp. 9912–9924, Curran Associates, Inc.

[19] D. Chicco, *Siamese Neural Networks: An Overview*, pp. 73–94, Springer US, New York, NY, 2021.

[20] J. Chen et al., "Dasnet: Dual attentive fully convolutional siamese networks for change detection in high-resolution satellite images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 1194–1206, 2021.

[21] S. Woo et al., "Cbam: Convolutional block attention module," in *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[22] Mustapha A. et al., "A deep dive into deep cluster," *CoRR*, vol. abs/2207.11839, 2022.

[23] Iris de Gélis, Thomas Corpetti, and Sébastien Lefèvre, "Change detection needs change information: improving deep 3d point cloud change detection," *arXiv preprint arXiv:2304.12639*, 2023.

[24] I. de Gélis et al., "Siamese kpconv: 3d multiple change detection from raw point clouds using deep learning," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 197, pp. 274–291, 2023.